BioTrHMM:基于迁移学习的生物医学命名实体识别算法 *

高冰涛，张　阳†，刘　斌

(西北农林科技大学 信息工程学院, 陕西 杨凌 712100)

摘　要： 传统的生物医学命名实体识别方法需要大量目标领域的标注数据，但是标注数据代价高昂。为了降低生物医学文本中命名实体识别对目标领域标注数据的需求，将生物医学文本中的命名实体识别问题化为基于迁移学习的隐马尔可夫模型问题。对要进行命名实体识别的目标领域数据集无须进行大量数据标注，通过迁移学习的方法实现对目标领域的识别分类。以相关领域数据为辅助数据集，利用数据引力的方法评估辅助数据集的样本在目标领域学习中的贡献程度，在辅助数据集和目标领域数据集上计算权值进行迁移学习。基于权值学习模型，构建基于迁移学习的隐马尔可夫模型算法 BioTrHMM。在 GENIA 语料库的数据集上的实验表明，BioTrHMM 算法比传统的隐马尔可夫模型算法具有更好的性能；仅需要少量的目标领域标注数据，即可具有较好的命名实体识别性能。

关键词： 迁移学习；隐马尔可夫模型；命名实体识别；文本挖掘

中图分类号： TP301.6　　doi: 10.3969/j.issn.1001-3695.2017.07.0702

BioTrHMM：named entity recognition algorithm based on transfer learning in biomedical texts

Gao Bingtao, Zhang Yang†, Liu Bin

(*College of Information Engineering*, *Northwest A&F University*, *Yangling*, *ShaanXi* 712100, *China*)

Abstract: Traditional methods of biomedical named entity recognition（NER） require a large amount of labeled data in the target domain, but the cost of tagging data is expensive. In order to reduce the requirement of labeled data in target domain for NER, the problem of NER in biomedical texts is transformed into a hidden Markov model based on transfer learning. The data sets in the target domain for NER do not need a large amount of labeled data to learn a model for the task by transfer learning. With the help of labeled data in source data sets across a different but related domain, and use the method of data gravitation to evaluate the contribution of samples in the auxiliary data sets about learning a model for the target domain. And calculate the weights of the data from the source domain and the data from the target domain. And then construct the hidden Markov model algorithm(BioTrHMM) based on the transfer learning. The experiment results on GENIA corpus show the BioTrHMM algorithm has better performance than the traditional algorithm of hidden Markov model, only uses small amount of labeled data in target domain.

Key Words: Transfer learning; Hidden Markov Model; Named Entity Recognition; Text mining

0　引言

传统的生物医学命名实体识别方法往往需要使用大量标注数据集构建模型，从而保证模型的分类预测性能。但是在现实情况中，通常本文感兴趣的领域中可获得的已标注数据很少，缺乏足够大的训练集训练强壮的模型，并且人工标注的代价高昂。迁移学习可以从相关领域数据集中学习知识，辅助学习目标领域的知识，协助解决目标领域的学习问题。利用相关领域的数据，可以大大减少对目标领域已标注数据的需求量，节约标注数据的高昂成本。

本文利用基于实例的迁移学习方法对辅助数据集进行知识迁移，协助解决目标领域的学习问题。为了降低对目标领域已标注数据的需求，算法需要处理如下问题：a）如何在目标领域标注数据较少的情况下得到性能较好的预测模型；b）如何实现跨领域知识迁移，从而辅助目标任务进行学习。本文算法利用数据引力方法评估辅助数据集中样本对目标学习问题的贡献程度，进而对辅助数据集中样本赋予权值，提出了基于样本的迁移学习方法。本文通过修改隐马尔可夫模型的学习算法和分类方法，提出基于权值的隐马尔可夫模型--BioTrHMM 算法。

隐马尔可夫模型（HMM）在传统的生物医学命名实体识别

基金项目： 国家自然科学基金资助项目（61602388）；中央高校基本科研业务费专项资金资助项目（2452015193，2452015194，2452016081）

作者简介： 高冰涛（1991-），男，山东烟台人，硕士研究生，主要研究方向为数据挖掘、机器学习；张阳（1975-），男（通信作者），教授，博士，主要研究方向为数据挖掘、机器学习（18706819983@163.com）；刘斌（1981-），男，讲师，博士，主要研究方向为机器学习、并行计算.

中应用非常广泛。比如基于单词相似度平滑技术的 HMM 命名实体识别分类器[1]、PowerBioNE 生物命名实体识别系统[2]、Zhang 等人[3,4]也指出 HMM 在生物医学领域中进行命名实体识别的有效性等。这些方法为了获得良好的预测性能，需要大量已标注的样本作为训练数据集来构建预测模型。

迁移学习[5]作为一种解决跨领域知识学习问题的学习方法，在命名实体识别和软件故障预测等领域[6~10]都有很好的应用。特别是基于实例的迁移学习方法[11~14]，如 nearest neighbour(NN) filter 和 transfer naive Bayes(TNB)等，通过度量样本对构建模型的贡献大小给样本分配不同的权重，辅助数据样本与目标数据越相似，则对构建分类分析模型起到的作用越大，所以赋予的权值越大[15]。本文采用文献[12]中的数据引力模型来评估辅助数据集中样本数据与目标数据集中样本数据的相似性，计算数据集中每个样本的权值。

本文算法仅需要少量目标领域标注数据，对已有相关但不同标注数据的辅助数据集进行迁移学习，基于辅助数据集和目标领域标注数据集构建预测模型，识别目标领域数据集中的命名实体。本文在 GENIA 语料库上针对不同角度进行了多组实验，实验结果表明本文提出的 BioTrHMM 算法在大大减少人工标注样本开销的情况下，比传统的 HMM 算法具有更好的预测性能。

1　问题定义

对于生物医学文本命名实体识别中需要大量的已标注样本，而人力标注开销大的问题，本文针对目标数据集 D_t，以相关但不同领域的数据集 D_s 为辅助数据集，将其转换为迁移学习场景下的 HMM 问题。给定训练数据集 D_l，$D_l = D_s \cup D_t$。在目标数据集上，$V = (v_1, v_2, ..., v_m)$ 为观测序列，$I = (i_1, i_2, ..., i_m)$ 为 V 对应的词性状态序列。本文的目标是通过对 D_l 中的样本赋予权值，完成对 D_s 的知识迁移，得到 $D_l' = (sample, w)$，其中，$sample$ 为样本，w 为样本对应的权值。在 D_l' 上构建一个 HMM 模型 f，使得对于给定 V，$V \in D_{t\text{-}test}$，$D_{t\text{-}test}$ 为与 D_t 同分布的目标测试集，有

$$f(V) \to I \tag{1}$$

即对于给定的观测序列 V，通过模型 $f(V)$ 对序列的词性状态进行识别分类，得到该序列所对应的词性状态序列 I，输出词性状态为实体类型的观测样本，完成命名实体识别。

2　BioTrHMM 算法

本文提出基于迁移学习的隐马尔可夫模型 BioTrHMM 算法，在使用较少的目标领域数据的情况下，基于目标数据集和辅助数据集构建模型，对目标数据集进行预测，依然具有较好的性能。本文通过评估辅助数据集中样本对目标学习问题的贡献程度，利用数据引力的方法对样本赋予权值，进行知识迁移，通过修改隐马尔可夫模型的学习算法，得到迁移学习场景下的隐马尔可夫模型。本文的技术路线可分为 4 个主要步骤：数据集的构建、知识的迁移学习、模型的学习以及预测与评估，如图 1 所示。其中，图 1 中的第 2 部分和第 3 部分构成本文的 BioTrHMM 算法，将在下文进行介绍。

2.1　基于实例的数据迁移

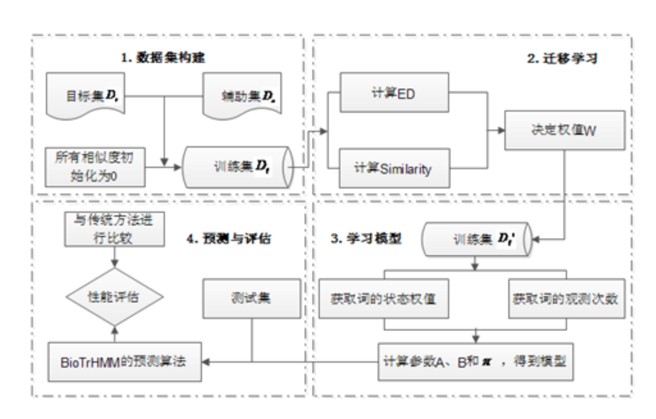


图 1　BioTrHMM 算法模块关系

为了给数据样本赋予权值，本文以辅助数据样本与目标数据样本之间的相似度评估辅助数据集对目标学习问题的贡献程度。

2.1.1 计算相似度

根据 D_t 中的词性和结构信息，计算 D_s 中的样本与 D_t 中样本的相似度。本文分别使用单词相似性和编辑距离对 D_s 中的样本数据和 D_t 中的样本数据的相似度进行计算。

定义 1　单词相似性（*Similarity*）是指两个不同的单词字符串中最大相同字符串的长度。

公式定义如下：

$$Similarity = \frac{l}{\max l} \tag{2}$$

其中：l 表示两个单词最大相同字符串长度，$\max l$ 表示两个单词中长单词的字符串长度。

编辑距离（edit distance）是指对于两个字符串，由一个转换成另一个所需的最少编辑操作次数。许可的编辑操作包括将一个字符替换成另一个字符，插入一个字符，删除一个字符。假设字符串 c 和 d 的长度分别为 y 和 e，则字符串 c 和 d 的编辑距离 $ED(e, y)$ 的计算公式[18]为

$$\begin{aligned} &ED(0,0) = 0 \\ &ED(0,y) = ED(y,0) = y \\ &ED(e,y) = \\ &\begin{cases} ED(e-1, y-1) & c_e = d_y \\ 1 + \min(ED(e-1, y), \\ \quad ED(e, y-1), & c_e \neq d_y \\ \quad ED(e-1, y-1)) \end{cases} \end{aligned} \tag{3}$$

2.1.2 计算权值

因为命名实体数据的词性类别属于 NN（noun, singular or mass：名词，单数；物资名词(不可数名词)）[16]，所以计算 D_s 中的词性为 NN 的样本数据与 D_t 中实体类型样本数据和词性为 NN 的样本数据的相似性。设 $Similarity_p$ 或者 $EditDistance_p$ 为 D_s 中样本与 D_t 中第 p 个实体类型样本或词性

为 NN 的样本的相似度，则对于目标样本的权值 W 可由式（4）或式（5）计算得到。其中，m_1 和 m_2 分别为两物体的质量,K 为常数。

$$W_p = G\frac{m_1 m_2}{r^2} = G\frac{K m_1 m_2 Similarity_p}{(2 - Similarity_p)^2} \propto \frac{Similarity_p}{(2 - Similarity_p)^2} \tag{4}$$

$$W_p = G\frac{m_1 m_2}{r^2} = G\frac{K m_1 m_2}{(1 + EditDistance_p)^2} \propto \frac{1}{(1 + EditDistance_p)^2} \tag{5}$$

D_s 中每个词性为 NN 的数据的最终权值为

$$W = \max(W_p) \qquad p = 1,2,\ldots,m \tag{6}$$

其中：m 为 D_t 中实体类型样本数与词性为 NN 的样本数之和。

通过上述方法对 D_l 中样本赋予权值，得到 D_l'，并用以构建模型。

2.2　BioTrHMM 的学习算法

本文使用的基本模型是隐马尔可夫模型，模型参数包括：状态转换概率矩阵 A、观测概率矩阵 B 和初始状态概率向量 π。

其中 $A = [a_{ij}]_{n\times n}$,其中 a_{ij} 表示在时刻 t 处于状态 q_i 的条件下在时刻 t+1 转移到状态 q_j 的概率 $P(i_{t+1} = q_j | i_t = q_i)$，$i = 1,2,\ldots,n$；$j = 1,2,\ldots,n$。其中，$Q = \{q_1, q_2, \ldots, q_n\}$ 是所有可能的状态的集合，n 为所有可能的状态数。本文对隐马尔可夫模型的参数 A 和 π 进行了修改。传统的模型参数学习方法是使用转换状态的次数计算得到状态转换概率，本文是在迁移学习场景下进行模型参数的学习，故本文中参数 A 的计算方式如下：

$$a_{ij} = \frac{w_{ij}}{w_i} = \frac{w_{ij}}{\sum_{j=1}^{n} w_{ij}} \quad i = 1,2,\ldots,n \tag{7}$$

其中：w_{ij} 表示状态 i 转移到状态 j 的权值，w_i 表示状态 i 发生转移的权值之和。

$B = [b_j(k)]_{n\times m}$,其中 $b_j(k)$ 表示在时刻 t 处于状态 q_j 的条件下生产观测值 v_k 的概率 $P(o_t = v_k | i_t = q_j)$，其中 $k = 1,2,\ldots,m$；$j = 1,2,\ldots,n$。

$$b_j(k) = \frac{N_{jk}}{N_j} \tag{8}$$

其中：N_{jk} 表示状态 j 的时候观测到 v_k 的次数，N_j 表示状态 j 的时候可能观测到的所以观测值的次数总和。

$\pi = (\pi_i)$, π_i 是时刻 $t = 1$ 处于状态 q_i 的概率 $P(i_1 = q_i)$,其中 $i = 1,2,\ldots,n$。

$$\pi_i = \frac{w_i}{\sum_{i=1}^{n} w_i} \quad i = 1,2,\ldots,n \tag{9}$$

其中，w_i 表示初始状态为 i 的权值。

在 D_L' 上通过统计的方式学习得到 A、B、π 三个参数，得到模型 f。

2.3　BioTrHMM 的分类算法

在 D_l' 上学习得到模型 f 后，本文进一步对维特比算法[13]进行了修改，并使用修改后的维特比算法进行分类分析。本文在 D_l' 上通过样本的权值，计算得到基于权值的状态转换矩阵代替原维特比算法中使用状态出现次数计算状态转换矩阵。在给定模型 f 的情况下，维特比算法可以有效地得到观测序列对应的状态序列，即得到给定观测文本序列对应的词性序列，通过观察对比得到的词性序列，本文可以得到命名实体类型的观测文本，达到命名实体识别的目的。下面给出本文中的预测算法：

在这里定义两个变量 δ 和 ψ，定义在时刻 t 状态为 i 的所有单个路径 $(i_1, i_2, \ldots, i_t)$ 中概率最大值为

$$\delta_t(i) = \max_{i_1,\ldots,i_{t-1}} P(i_t = i, i_{t-1}, \ldots, i_1, v_t, \ldots, v_1 | \lambda) = \max_{1\le j\le N}[\delta_t(j) a_{ji}] b_i(v_{t+1}) \quad i = 1,2,\ldots,N; t = 1,2,\ldots,T-1 \tag{10}$$

定义在时刻 t 状态为 i 的所有单个路径 $(i_1, i_2, \ldots, i_{t-1}, i_t)$ 中概率最大的路径的第 t-1 个节点为

$$\psi_t(i) = \arg\max_{1\le j\le N}[\delta_{t-1}(j) a_{ji}], i = 1,2,\ldots,N \tag{11}$$

具体预测如下:

a）初始化。

$$\delta_1(i) = \pi_i b_i(v_1) = \frac{w_i}{\sum_{i=1}^{n} w_i} b_i(v_1), \quad i = 1,2,\ldots,n; \psi_1(i) = 0, i = 1,2,\ldots,n \tag{12}$$

首先对与给定观测序列，在时刻 t=1 时计算得到状态为 i 的所有单个路径 (i_1) 中概率最大的路径。由于时刻 t=1 没有前一时刻，所以 $\psi_1(i) = 0, i = 1,2,\ldots,n$。

b）递推。对，$t = 2,3,\ldots,T$

$$\delta_t(i) = \max_{1\le j\le n}[\delta_{t-1}(j) a_{ji}] b_i(v_t) = \max_{1\le j\le n}[\delta_{t-1}(j)\frac{w_{ji}}{\sum_{i=1}^{n} w_{ji}}] b_i(v_t), i = 1,2,\ldots,n \tag{13}$$

$$\psi_t(i) = \arg\max_{1\le j\le n}[\delta_{t-1}(j) a_{ji}] = \arg\max_{1\le j\le n}[\delta_{t-1}(j)\frac{w_{ji}}{\sum_{i=1}^{n} w_{ji}}], i = 1,2,\ldots,n \tag{14}$$

对时刻 $t = 2,3,\ldots,T$ 依次计算得到出现的状态路径中概率最

大的路径，并得到概率最大路径的第 t-1 个节点。

c）终止。

$$P^* = \max_{1\le i\le N} \delta_T(i) \\ i_T^* = \arg\max_{1\le i\le N}[\delta_T(i)] \tag{15}$$

当时刻 $t=T$ 时，可以计算得到最后一个时刻概率最大路径对应的节点和概率最大路径中前一时刻对应的节点。

d）最优路径回溯，对 $t=T-1, T-2, \ldots, 1$

$$i_t^* = \psi_{t+1}(i_{t+1}^*) \tag{16}$$

由时刻 $t=T$ 依次回溯到 $t=1$ 时刻，就可以得到观测序列对应的概率最大路径，即最优路径。最终求得的最优路径，也就是词所对应的词性状态，根据得到的状态序列，输出命名实体类型对应的观测样本，完成命名实体识别。

表 1　GENIA V3.02 语料库中实体标签分布

Protein/%	DNA/%	RNA/%	Cell type /%	Cell line /%	No-entity /%
10.45	3.95	0.40	4.00	1.90	79.30

表 2　辅助集中实体标签分布

DNA/%	RNA/%	Cell type/%	Cell line/%	No-entity/%
3.95	0.40	4.00	1.90	89.75

3　实验及结果分析

为了验证 BioTrHMM 算法的性能，本文在 GENIA v3.02 语料库上进行了实验。

3.1　实验设置

为了验证本文的算法对生物医学命名实体识别的预测性能，选取传统的 HMM 算法与本文提出的基于迁移学习的 BioTrHMM 算法进行比较。目前，最常用的生物医学标注语料库是 GENIA v3.02 语料库，该语料库包含了来自 MEDLINE 的 2 000 个摘要标注文本（约 360 000 个单词），并且包含 36 个词性类别，其中包含 5 个生物医学实体类型。本文使用了 GENIA v3.02 语料库（http://www.nactem.ac.uk/genia/genia- corpus）的数据进行了实验。本文识别的是蛋白质命名实体，采用了精确率、召回率和 F 值[17]作为评价指标。GENIA v3.02 语料库中实体标签分布说明如表 1 所示。

本文中 D_t 是含有蛋白质命名实体标签和其他词性标签的目标集，D_s 是把蛋白质命名实体标签处理为 NN 类型的辅助集，辅助集中标签分布如表 2 所示。

本文中设置了三个参数 α，β，γ。其中 $\alpha|D_t|$ 表示目标集的大小；$\beta|D_s|$ 表示辅助集的大小；γ 表示所用数据集所占 GENIA 语料库的比例，当使用全部 GENIA v3. 02 语料库时，γ 值为 1。本文通过对每组实验进行十折交叉验证的方法，确保结果的有效性。

3.2　实验结果

为了验证 BioTrHMM 算法的性能，本文分别从不同角度进行了实验。实验如下：

3.2.1 针对 α 的实验

本文针对不同的 α 取值分别对 BioTrHMM 和 HMM 进行了实验，实验结果如图 2 所示，实验结果表明同样大小的 D_t 下，通过对 D_s 中知识的迁移学习，BioTrHMM 算法的识别性能显著优于 HMM 的识别性能。

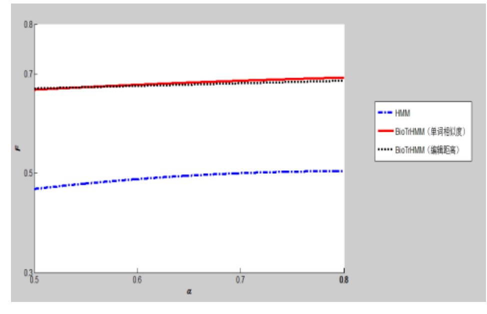


图 2　针对 α 的实验

通过对辅助集的学习，BioTrHMM 学到了更多对目标任务有贡献的知识，使得学习得到的模型更加健壮。BioTrHMM 与 HMM 有相同的目标集，但是 BioTrHMM 通过使用现有分布不同的辅助集在不增加人工标注成本的情况下，大大提升了算法的性能。

3.2.2 针对 β/α 的实验

本文为了研究 D_t 大小对算法性能的影响，在训练集大小相同情况下，针对不同 β/α 进行了实验。本文把 D_L 中 β 和 α 的数据集大小比例设定为 2：1、3：1、4：1 和 5:1。表 3 是 $\gamma=1$ 时的实验结果对比。

尽管随着 D_t 数量的减少 BioTrHMM 算法的性能有所降低，但是仍然与传统的隐马尔可夫模型的预测效果相当。

表 3　BioTrHMM 与 HMM 实验结果对比

方法	β/α	精确率	召回率	F 值
HMM	-	0.6355	0.5399	0.5797
BioTrHMM（Similarity）	2:1	**0.8955**	0.5201	**0.6519**
	3:1	0.8695	0.4974	0.6226
	4:1	0.8684	0.4701	0.6009
	5:1	0.8822	0.4578	**0.5848**
BioTrHMM (EditDistance)	2:1	**0.8761**	**0.5363**	**0.6541**
	3:1	0.8704	0.4980	0.6233
	4:1	0.8677	0.4723	0.6026
	5:1	0.8697	0.4453	**0.5811**

3.2.3 针对 γ 的实验

为了进一步探讨算法的有效性，本文还分别在不同大小的数据集上进行了实验。本文分别在 $\gamma=0.8$ 和 $\gamma=0.6$ 是进行了实验，实验结果如表 4 和 5 所示。实验结果表明，标注的目标数据集的规模对传统的 HMM 算法的性能具有较大的影响。尽管数据集规模有所减小，但是 BioTrHMM 算法依然具有良好的预

测分类性能。同时，BioTrHMM 算法在保证分类分析性能的前提下，有效的降低了对标注目标数据的需求量，减少了人工标注数据的开销。

综上实验结果表明本文提出的 BioTrHMM 算法通过对跨领域知识的迁移，可以在生物医学蛋白质命名实体标注数据较少的情况下，可以达到较好的分类分析效果。此外，以上实验结果表明相似度度量方法的不同对本文提出的 BioTrHMM 算法几乎没用影响。使用不同的相似度度量方法，BioTrHMM 算法都可以实现比传统隐马尔可夫模型更好的预测效果。由于 HMM 与 BioTrHMM 算法实验使用的训练集和测试集的大小都相同，而 BioTrHMM 算法算法中训练集是由两部分组成，故在 BioTrHMM 算法实验中目标集的大小远远小于传统隐马尔可夫模型算法实验中训练集的大小。因此实验结果表明与传统隐马尔可夫模型相比，BioTrHMM 算法仅使用其 1/3，甚至更少的标注目标集，就可以达到良好的分类分析效果。

表 4　$\gamma = 0.8$ 实验结果对比

方法	β/α	精确率	召回率	F 值
HMM	-	0.5591	0.4561	0.5024
BioTrHMM（Similarity）	2:1	**0.8909**	0.4298	**0.5799**
	3:1	0.9508	0.3919	0.5550
	4:1	0.9828	0.3851	0.5534
	5:1	**0.9837**	0.3846	**0.5530**
BioTrHMM（EditDistance）	2:1	**0.8909**	0.4298	**0.5799**
	3:1	0.9677	0.4054	0.5714
	4:1	0.9661	0.3851	0.5507
	5:1	**0.9716**	0.3829	**0.5493**

表 5　$\gamma = 0.6$ 实验结果对比

方法	β/α	精确率	召回率	F 值
HMM	-	0.5205	0.4524	0.4841
BioTrHMM（Similarity）	2:1	**0.8780**	0.4286	**0.5760**
	3:1	0.9712	0.4052	0.5731
	4:1	0.9778	0.3793	0.5466
	5:1	**0.9837**	0.3707	**0.5385**
BioTrHMM（EditDistance）	2:1	**0.8810**	0.4405	**0.5873**
	3:1	0.9796	0.4138	0.5818
	4:1	0.9565	0.3793	0.5432
	5:1	**0.9773**	0.3707	**0.5375**

4　结束语

本文针对传统生物医学文本的命名实体识别方法需要大量的标注目标样本，但是现实中标注样本困难的问题，提出了基于迁移学习的隐马尔可夫算法 BioTrHMM。算法通过计算辅助集数据集和目标数据集集中样本数据的相似度，根据数据对目标任务贡献的大小，赋予辅助集数据集中样本数据权值，实现基于实例的迁移学习。本文通过对 HMM 学习算法进行改进，在加权数据集中学习 HMM 的模型参数，建立迁移学习条件下的预测模型。实验结果表明，BioTrHMM 在使用较少目标领域已标注数据的情况下，具有更好的预测性能。本文提出的方法不仅可以用于生物医学文本的命名实体识别中，同时可以推广到文本挖掘的命名实体识别当中。

在部分本文仅对 HMM 算法进行了改进，很多研究表明条件随机场在命名实体识别中比 HMM 具有更好的识别性能[19,20]。基于此研究成果和研究现状，未来工作考虑将在条件随机场基础上对命名实体识别进行迁移学习，从而提升对命名实体的识别性能。

参考文献：

[1] Horn H, Schoof E M, Kim J, et al, KinomeXplorer: an integrated platform for kinome biology studies [J]. Nature Methods, 2014, 11 (6) , 603–604.

[2] Srinivasagan K G, Suganthi S, Jeyashenbagavalli N. NER for Hindi language using association rules [C]// Proc of International Conference on Data Mining and Intelligent Computing. 2014.

[3] Gayen V, Sarkar K. An HMM based named entity recognition system for Indian languages: The JU System at ICON [R]. Published in ArXiv, 2013.

[4] Arnold A, Nallapati R, Cohen W W, et al. Exploiting feature hierarchy for transfer learning in named entity recognition [C]// Proc of the 46th Annual Meeting of the Association for Computational Linguistics. 2008, 245-253.

[5] Pan Jialin, Yang Qiang. A survey on transfer learning [J] , IEEE Trans on Knowledge and Data Englineeging, 2010, 22 (10): 1345-1359.

[6] Liu Jie, Yu Kai, Zhang Yi, et al, Training conditional random field using transfer learning for gesture recognition [C]// Proc of IEEE International Conference on Data Mining. 2010.

[7] Magimai-Doss M, Rasipuram, Aradilla G, et al. Grapheme-based automatic speech recognition using KL-HMM [C]// Proc of InterSpeech. 2011.

[8] Cui Xiaodong, Huang Jing, Chien J T. Multi-view and multi-objective semi-supervised learning for HMM-based automatic speech recognition [J]. IEEE Trans on Audio, Speech, and Language Processing, 2012, 20 (7): 1923-1935.

[9] Hu Hao, Zheng Wenchen, Yang Qiang, Cross-domain activity recognition via transfer learning [J]. Pervasive and Mobile Computing, 2011, 7 (3): 344–358.

[10] Cook D, Feuz K D, Narayanan C, et al, Transfer learning for activity recognition: a survey [J]. Knowledge of Information System. 2010, 36: 537–556.

[11] Pan Jialin, Toh Zhiqiang et al. Transfer joint embedding for cross-domain named entity recognition [J]. ACM Trans on Informaition Systems, 2013, 31 (2): Article 7.

[12] Ma Ying, Luo Guangchun, Zeng Xue, et al, Transfer learning for cross-company software defect prediction [J]. Information and Software Technology, 2012, 54: 248-256.

[13] Rabiner L, Juang B, An introduction to hidden Markov models [J]. IEEE

ASSP Magazine, 1986.

[14] 庄福振，罗平，何清，等，迁移学习研究进展 [J]. 软件学报, 2015, 26 (1): 26-39.].

[15] Bui Q C, Katrenko S, Sloot P M A. A hybrid approach to extract protein-protein interactions [J]. Bioinformatics, 2011, 27 (2): 259-265.

[16] Huang Yuanhua, Xu Bosen, Zhou Xueya, et al. Systematic characterization and prediction of post-translational modification cross-talk [J] , Molecular & Cellular Proteomics, 2015, 14 (3): 761-770.

[17] 张阳，李建良，胡正国. NewsGrouper: 一个自动抽取重要新闻的软件工具, 计算机工程, 2002, 28 (4): 83-84.

[18] Teixeira, J, Sarmento L, Oliveira E. A bootstrapping approach for training a NER with conditional random fields [C]// Portuguese Conference on Artificial Intelligence. Berlin: Springer, 2011.

[19] Konkol M, Konop´ık M. CRF-based Czech named entity recognizer and consolidation of Czech NER research [J]. Berlin: Springer-Verlag, 2013: 153–160.